图书情报工作
第 65 卷 第 23 期 2021 年 12 月



# 融合 Word2vec 和 WGRA 的社会化问答社区答案有用性排序方法研究——以携程问答为例*

■ 郭顺利 步辉
曲阜师范大学传媒学院 日照 276800

摘 要:[目的/意义]为解决社会化问答社区用户信息需求多样化和答案冗余过载问题,提出面向用户个性化需求的答案有用性排序方法,协助用户高效筛选和获取有用的答案知识。[方法/过程]首先通过文献调研和专家咨询法,从答案特征、回答者特征、答案的时效性 3 个维度构建答案有用性评价指标体系;然后,从语义层面融合用户个性化需求,设计融合加权灰色关联分析法和 Word2vec 的答案有用性排序方法,实现面向用户需求的答案排序。[结果/结论]通过实验结果的对比分析发现与基于"点赞数"和"回答时间"等传统的排序方法相比,笔者设计的答案有用性排序方法的用户满意度更高,更能够满足用户的个性化知识需求。

关键词:用户需求 答案有用性 加权灰色关联度 Word2vec 社会化问答社区
分类号:G252
DOI:10.13266/j.issn.0252-3116.2021.23.014



## 1 引言

近年来,随着社会化问答社区的发展和推广,社会化问答社区已逐渐成为网络用户获取高质量信息或专业知识的重要途径,并且开始向更高质量、专业化、社会化的方向发展。然而,由于社会化问答社区的问题和答案以用户生成为主,用户信息素养参差不齐,社区监管力度不足等原因,导致网络问答社区的用户生成内容出现冗余繁杂、质量良莠不齐,以及答案与用户需求契合度不高等问题,更是没有实现面向用户需求的个性化排序。因此,如何在大量的答案文本中识别出匹配用户需求、有用性较高的个性化答案成为社会化问答社区亟需解决的问题。该问题的解决对于提升社会化问答社区服务质量和用户粘性具有极为重要的意义。

目前,国内外针对社会化问答社区答案方面的研究主要包括答案推荐[1]、答案质量评价[2]、答案排序[3]、答案融合[4]等。其中,答案推荐是针对提问者而言,通过一些推荐算法对问题的候选答案作自动排序,从而使提问者能更快捷地选择最佳答案,针对答案推荐方面的研究主要集中于采用不同角度和方法识别最佳答案。冯文政等通过双向 LSTM、词向量、2D 神经网络等深度学习模型与 TF-IDF、LCS 等传统特征结合来筛选最佳答案[5];谢正文等则是将思路转变为通过寻找两者间更细粒度的语义信息筛选最优答案[6];W. Ma 等利用长短时记忆网络和卷积神经网络提取问答对的语义特征,计算问题与答案之间的匹配度,从而实现对答案的推荐[7]。而笔者提出的答案有用性排序则是针对大多数浏览者而言的,即根据用户个人所感知的答案有用性来个性化地为用户呈现答案顺序。可见,答案推荐与答案有用性排序存在本质上区别。另外,社会化问答社区答案有用性与答案质量既有重合之处又有所不同。答案质量,通俗来讲就是答案的好坏,好坏的评判一般是基于某种标准,那么答案质量的评判就是基于答案的特征人为给定标准进行评判。目前,针对答案质量方面的研究主要集中于从答案特征角度进行答案质量评价。J. Jiwoon 等[8]首次提出采用非文本特征即答案的长度、采纳率、推荐次数及页面点击率等,并利用最大熵模型成功进行了答案质量识别;

* 本文系国家社会科学青年基金项目"基于认知计算的网络问答社区知识的深度聚合及精准服务研究"(项目编号:20CTQ028)研究成果之一。
作者简介:郭顺利,讲师,博士,E-mail:guosl777@163.com;步辉,硕士研究生。
收稿日期:2021-06-23 修回日期:2021-09-22 本文起止页码:126-135 本文责任编辑:徐健

E. Agichtein[9]等则创新融合了非文本特征和文本特征，利用 C4.5 决策树全面化分析了 Yahoo! Answers 的答案质量。然而答案有用性的概念则是来自于信息接受模型，根据信息接受模型可知，用户接收外界信息时将信息质量和信息源的可信度作为信息有用性的评判标准[10]。因此，社会化问答社区答案有用性是指用户在使用社会化问答社区检索或者浏览答案时根据个人信息需求所感知的答案价值，为用户在解决问题时提供帮助的程度。由此可知，答案有用性排序是指基于用户所感知到的答案价值和有用程度实现的面向用户需求的个性化排序结果。当前国内外学者针对问答社区答案有用性研究主要从以下两个方面展开：

（1）在线问答社区答案有用性的影响因素方面研究。许多学者基于不同的理论，分别从不同角度验证了各类因素对答案有用性的影响。S. M. Mudambi 等以亚马逊网站为研究对象，发现在线评论有用性与评论深度、评论情感极性以及商品类型有关[11]；谢陈博从信息接受理论出发构建答案有用性理论模型，并通过实证研究证明外向中心度对答案的有用性影响不显著，内向中心度对答案有用性影响显著且影响程度最大[12]；曾珍妮认为历史提问经验对答案有用性没有显著影响，答案长度、情感倾向、使用的图片数量都对答案有用性存在显著的正向影响，答者的历史回答经验、发表文章的经验、内向网络中心度、外向网络中心度这 4 个因素都对用户的感知有用性具有正向作用[13]；王晨指出回答的文本长度、经过再次编辑、文本专业性对于回答有用性有显著的正向影响，回答中使用的图片数量和情感倾向对于回答有用性的影响是负向的。回答用户的历史提问数量和回答数量、用户取得知乎认证身份、用户的内向网络中心度和外向网络中心度对于回答的有用性也存在正向影响，而专栏数量的影响是负向的[14]。

（2）社会化问答社区答案排序方法方面的研究。学者们基于不同的理论基础，采用不同研究方法开展答案排序研究。C. Shah 等从相关性、信息量、完整性等维度对 Yahoo! Answer 的答案进行人工评分，探究答案有用程度[15]；李晨等提取答案的文本和非文本特征，采用人工标注和逻辑回归的方法对数据集进行质量分类[16]；Z. M. Zhou 等将用户信息融入 SVMRank、List-Net 排序模型，排序结果更有优越性[17]；来社安和蔡中民从语义相似度角度出发，计算问题和答案的相似度和权值并加以调整，从而选出最佳答案[18]；易明和张婷婷认为对答案质量指标体系利用 K-Medois 聚类算法和粗糙集理论修正后，运用加权灰色关联分析法计算灰色关联度产生的排序结果的用户满意度更高[19]；刘瑜和袁健对 TEM 模型（Tree-enhanced Embedding Model）加以改进，分析用户行为，形成新的答案排序和自动筛选模型；L. Yang 等基于 TEM 模型结合文本内容模型和链接结构分析进行建模，通过 CQARank 对 Stack Overflow 进行实证研究，计算出答案的主题相似性和用户权威性，进而产生排序结果[20]。

通过梳理已有研究发现，问答社区答案质量或有用性方面研究受到国内外学者们的关注，产生了一系列的研究成果。利用不同理论、多种角度分析答案有用性的影响因素和各指标的影响结果，并积极改进排序方法，为本文的研究奠定了一定的理论基础和参考依据。然而，已有研究主要是为了识别并评判答案质量，少有将答案有用性和答案排序结合进行研究，更没有深入到答案的语义层面考虑用户的个性化需求。用户通过检索或者浏览等方式获取有用答案的过程有一定的时间忍耐度，更期待能够快速搜寻到匹配自身需求答案，趋向于消耗最低的成本获得最佳答案。基于此，笔者借鉴已有相关研究，深入答案语义层面提出一种面向用户需求的答案有用性排序方法。首先从答案特征、回答者特征和答案时效性 3 个维度筛选出影响社会化问答社区答案有用性关键性指标并分别对其进行量化；然后结合使用熵权法、加权灰色关联分析法、Word2vec 等方法提出了社会化问答社区答案有用性排序的新方法。最后，选取携程旅游网问答社区的杭州话题为研究对象，验证笔者提出的答案有用性排序方法的有效性和科学性。

## 2 社会化问答社区答案有用性评价指标体系构建

### 2.1 关键评价指标的选取与量化

本研究基于笔者已发表论文《社会化问答社区用户生成答案质量自动化评价研究－以“知乎”为例》[21]，认为用户在评价答案有用性过程中受到多方面因素的影响。一般情况下需要考虑答案文本内容质量、回答者质量、时效性等维度因素，大部分研究也证实了这 3 类特征对答案有用性产生影响。易明等通过 11 种学习算法比较得出点赞数、粉丝数对于答案质量的影响程度最大[22]；施国良利用内容分析法和回归模型检验发现答题者的影响力、答案及时性、答案长度对于答案认可都具有正向影响[23]；翟倩认为产品的属性

特征词和情感表达能够提高用户浏览时对在线评论的可信度和感知有用性[24]。同时,笔者又查阅了已有研究文献以及向有关专家咨询,从答案内容特征、回答者特征和答案时效性 3 个角度确立了答案有用性评价指标。具体指标及其量化方法,如表 1 所示:

**表 1 社会化问答社区答案有用性排序指标与量化方法**

| 指标类型 | 指标名称 | 指标含义及作用 | 指标量化方法 | 主要文献来源 |
|---|---|---|---|---|
| 答案内容特征 | 文本长度 | 答案文本中所包含的字符数量。答案的文本长度与用户的感知有用性之间存在着倒 U 型关系。答案文本中的字符在一定字数内对于用户所感知的有用性是更高的 | 答案文本的有效字符数量总数,可由数据采集获得并通过 Excel 函数计算 | [25][13][26][16] |
| | 图片数量 | 答案中所包含的有效图片数量。答案中添加图片可以丰富答案内容,增加用户对答案的认同感,提高答案的可阅读性 | 答案中有效图片数量总数,可由数据采集获得 | [27][28][9][29] |
| | 答案评论 | 答案下方被评论数量。被评论次数越多的答案,其热度越高。用户可以在评论中获得除答案自身外的额外或补充知识 | 答案下方被评论的总数,可由数据采集获得 | [9][5] |
| | 属性描述词 | 答案中关于问题主体属性的描述词语。属性描述词直接体现答案与问题的匹配程度,一般来说,属性描述词越多,答案匹配度越高 | 答案中的关于问题的属性特征词的总数。首先爬取某目的地下的所有问答对并对其进切分词和去停用词,经过人工筛选判断形成语料库;将已经预处理过的答案与语料库进行一一对比并累积比对成功次数记做属性特征词总数 | [30][31] |
| | 情感分析 | 答案中表达正向情感倾向的值。答案的情感分析能够向用户传达答案中的情感偏好,影响用户对答案有用性的感知 | 答案中的情感分析值。通过 python 的 snownlp 库计算情感分析值 | [32][33][34][35] |
| | 答案获赞 | 答案所获得点赞数。答案的点赞数越高,用户对答案的认同度越高 | 答案所获的点赞数量总数,可由数据采集获得 | [36][28] |
| 回答者特征 | 回答者权威 | 回答者的平台影响力。经调研发现,回答者的粉丝数量越多,其平台影响力越大,其发表的答案质量越高的可能性越大 | 回答者的粉丝数量,可由数据采集获得 | [14][15][12] |
| | 回答者获赞 | 回答者的总体获赞数。一方面回答者的总体获赞数能表达出回答者在某些问题或某些领域内的专业程度;另一方面可以体现回答者对于问答社区的知识贡献程度 | 回答者所有答案获赞总数,可由数据采集获得 | [37][12] |
| 答案时效性 | 时效性 | 答案发布时间与答案被阅读的时间差值。问题提出时间与答案发布时间的差值越大,无论是对于提问者还是后期的浏览者来说,答案越不及时,答案对于用户的有用程度越低 | 答案发布时间与答案被阅读的时间差值,以天为计数单位,可由数据采集获得并通过 Excel 函数计算 | [38][39][40] |



### 2.2 基于熵权法的答案有用性指标权重赋值

答案有用性指标的权重分配与赋值对后续答案有用性的排序起着至关重要的作用。由于权重反映的是指标在整个答案的有用性排序指标体系中的重要性,关系到指标对于排序结果的贡献程度,必须科学合理地根据每个指标的重要程度赋予不同的权重。熵权法作为典型的指标权重赋权方法,具有广泛的应用。熵权法认为某项指标的熵值越大,其信息量越大,内容越丰富,该指标对用户的有用性越强,其所占权重也应较大[41]。熵值可以表示信息的有用程度,当信息的熵值达到最大,即熵值为零时,意味着信息的有用性也为零。熵权法根据各指标所提供的信息量来确定指标的具体权重,是一种相对客观的赋权法。学者们已经将熵权法广泛应用于指标权重赋值。例如:李帅等利用熵权法和层次分析法为宁夏城市人居环境质量评价确定指标权重[42];信桂新等将熵权法运用于高标准的基本农田建设后效应评价体系的构建中[43]。熵权法在各领域的指标赋值方面的应用十分广泛,也得到了大家的一致认可。熵权法的优势在于一方面相较于其他指标赋权方法对于有用信息的筛选更加准确;另一方面熵权法能够排除传统赋权法由于人为主观性太强对实验结果产生的负面影响,增加权重的科学性和可信性。因此,笔者采用熵权法对答案有用性指标进行赋权,根据实际方法产生的数据结果确定答案有用性指标的权重。根据熵权法的内在原理,熵权法的基本步骤如下[44]:

(1)指标数据标准化。为了防止各指标量纲不统一造成实验误差,需要先将各指标的原始数据进行标准化处理。假设有 K 条数据,每条数据有 X 项评价指标,标准化处理后的指标为 y。具体公式如下:

$$y_{ij}=\frac{x_{ij}-min(x_{ij})}{max(x_{ij})-min(x_{ij})}\qquad \text{公式(1)}$$

(2)计算指标的熵值。根据信息论中信息熵的定义,一条数据的信息熵的计算公式如下所示:

$$E=-\ln(n)^{-1}\sum_{i=1}^{n}p_{ij}\ln(p_{ij})(i=1,2,\cdots,m;j=1,2,\cdots,n)\qquad \text{公式(2)}$$

其中,$p_{ij}=y_{ij}\sum_{i=1}^{n}y_{ij}$,如果 $p_{ij}=0$,则定义 $\lim_{y_{ij}=0}p_{ij}\ln(p_{ij})=0$。

(3)确定各指标的权重。根据信息熵的计算公式,计算出各个指标的信息熵为 $E_1,E_2,E_3,\cdots,E_k$。接着通过信息熵计算各指标的权重 $W_i$,并且 $0\leqslant W_i\leqslant 1$, $\sum_{k=1}^{n}W_i=1$。

$$W_i=\frac{1-E_i}{k-\sum E_i}(i=1,2,\cdots,k)\qquad \text{公式(3)}$$

## 3 融合加权灰色关联分析法与 Word2vec 算法的有用性排序方法过程

### 3.1 相关技术方法介绍

#### 3.1.1 Word2vec 算法

Word2vec 词向量模型是由 Tomas Mikolov 首先提出,主要思想是利用空间向量来表示单词。Word2vec 是通过训练文本来将其转化为 K 维向量运算,利用向量在空间的相似性来表示文本在语义上的相似度。词语经过训练在空间位置中转化成点,每个点代表一个单词,通过测量空间中的词向量的距离得到词语之间的相似性。因此,笔者采用 Word2vec 计算问答对的语义相似度。

Word2vec 包含 CBOW 和 Skip-gram 两个模型。CBOW 是通过上下文来预测当前词语的概率;Skip-gram 是通过当前词语来预测上下文词语的概率。两种训练模型虽然方向相反,但原理相似而且本质上都是以 Huffman 树作为基础,构建一个多层神经网络,在给定文本中获取对应的输入与输出,通过不断地训练与修改参数,最后获得词向量。通过查阅文献得知 Skip-gram 模型在处理专业领域文本方面更加优越,因此笔者选择 Skip-gram 模型训练词向量。具体工作原理的模型架构见图 1。

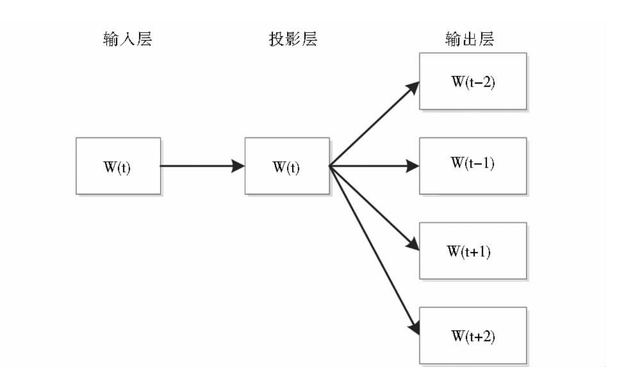



图 1 skip-gram 模型结构

#### 3.1.2 加权灰色关联分析法

灰色关联分析法[45-46]来自于我国学者邓聚龙在 1982 年提出的灰色系统理论,其基本思想是利用数学的方法表示各因素的数据,根据实验数据和参考数据曲线几何形状的拟合程度来判断灰色关联程度。一般地,在进行灰色关联分析时,通常采用各时点的灰色关联系数的算术平均数作为灰色关联度,这样没有考虑比较数列与参考数列里各元素的信息熵值,会造成一定程度上的信息损失,不能正确反映实验数据与参考序列的关系。因此,笔者运用基于熵权法优化的加权灰色关联分析法(Weighted Grey Relational Analysis, WGRA)来计算答案的加权灰色关联度。加权灰色关联分析法的基本步骤如下:

(1)确定分析数列,即分别确定参考数列和比较数列。将各条答案指标量化后的数据构成分析数列,设分析数列为 $X(i)=\{X(k)|k=1,2,\cdots,n\},i=1,2,\cdots,m$,$m$ 是每条问题下答案的具体条数。参考数列作为比较的标准,应该选取各指标的最优值,设参考数列为 $Y=\{Y(k)|k=1,2,\cdots,n\}$。

(2)无量纲化处理数据,因为各指标的初始量纲和数量级可能有所不同,为了准确地进行分析比较,将数据进行初值化或均值化处理。

(3)计算关联系数,即根据公式分别计算每个比较数列与参考数列对应元素的关联系数。

首先,计算比较序列与参考序列对应元素的绝对差值,计算公式如下:

$|x_o(k)-x_i|(k=1,2,\cdots,m;i=1,2,\cdots,n$;$n$ 为评价对象个数) 公式(4)

其次,计算 $\underset{i=1}{minn}$ 和 $\underset{i=1}{maxn}$,如公式(5)(6)所示:

$$\underset{i=1}{minn}=\underset{k=1}{minm}|x_0(k)-x_i(k)|\qquad \text{公式(5)}$$

$$\underset{i=1}{maxn}=\underset{k=1}{maxm}|x_0(k)-x_i(k)|\qquad \text{公式(6)}$$

接着,根据公式(7)计算关联系数 $\xi_i(k)$。

$$\xi_i(k)=\frac{\underset{i}{min}\,\underset{k}{min}|(x_0(k)-x_i(k)+\rho\cdot\underset{i}{max}\,\underset{k}{max}|x_0(k)-x_i(k)||}{|x_0(k)-x_i(k)|+\rho\cdot\underset{i}{max}\,\underset{k}{max}|x_0(k)-x_i(k)|}\qquad \text{公式(7)}$$

在式(7)中,ρ 为分辨系数,在(0,1)内取值,若 ρ 越小,关联系数间差异越大,区分能力越强。通常 ρ 取 0.5。

(4)计算加权关联度,即计算关联系数在各个时刻的平均值,用来表示比较数列与参考数列的具体关联程度。笔者运用熵权法计算出的指标权重 $W_i$ 优化灰色关联分析法形成加权灰色关联度方法,其计算公式(8)如下:

$$\gamma_i = \sum_{k=1}^{n} \gamma_i(k)\ W_i, k=1,2,\cdots,n \quad 公式(8)$$

### 3.2 融合 Word2vec 和灰色关联分析的答案有用性排序方法步骤

社会化问答社区用户的需求和其他环境下用户需求有所不同,用户在其个人知识需求的原动力驱动下将会产生一系列的知识获取行为,其中用户通过问答社区进行提问、回答或浏览相关问题及答案是重要的信息获取途径之一[47]。因此,社会化问答社区里的问答对不仅可为潜在用户提供决策参考,更重要的是开发者可从中挖掘出用户需求[48]。值得一提的是 Word2vec 不仅可以将文本数据转化为便于处理的数值型数据,而且擅长深入语义层面挖掘用户潜在需求信息。至于加权灰色关联分析法,一方面结合熵权法根据信息的丰富程度为其赋权,另一方面根据灰色关联分析法的原理识别出每条答案与标准答案的接近程度。以往研究往往聚焦于排序方法的创新,忽略了用户需求这一重要因素。综上所述,笔者选择融合 WGRA 和 Word2vec 算法,综合考虑社会化问答社区答案的特点以及出于体现大多数用户的多样化需求。

融合 WGRA 与 Word2vec 算法的答案有用性排序方法主要包括 3 个部分:首先构建答案有用性指标体系;然后根据熵权法确定各指标的具体权重;然后分别根据灰色关联分析法结合权重和 Word2vec 算法计算出答案的加权灰色关联度和问答对,最后实现社会化问答社区的答案排序并进行实验结果的对比分析。答案有用性排序的具体实现步骤,如图 2 所示:

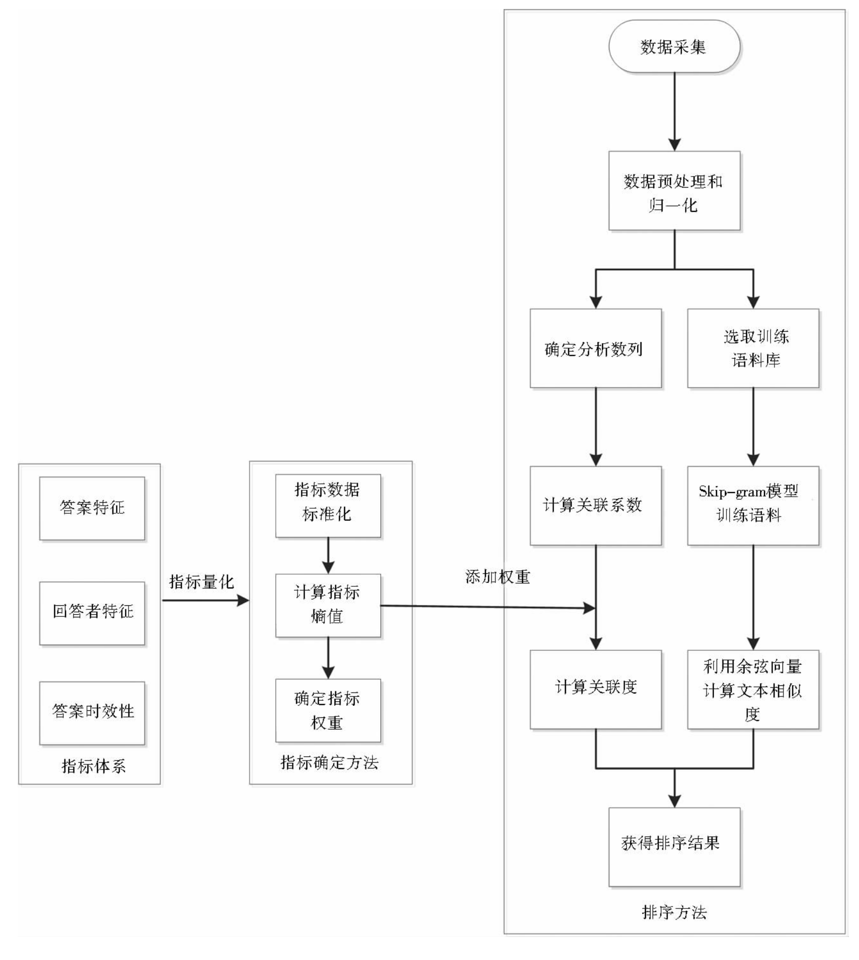



图 2 社会化问答社区答案有用性排序方法流程

根据用户的信息需求状态之间的关系可知,客观的信息需求通过认知内化成意识到的信息需求进而通过提问外化成表达出的信息需求[49]。因此,笔者以用户会通过提出的问题来充分而准确地表达其信息需求

为出发点进行算法流程设计，具体如下：

Step1：数据预处理和标准化。将采集到的数据通过 python 自编程序进行切词、分词，删除无用信息，进行标准化处理。

Step2：将各项指标按照具体量化方法转化成可处理的数值数据，接着通过熵权法计算指标的具体权重 $W_i$，进而结合灰色关联分析法计算加权灰色关联度 $\gamma_i$。

Step3：输入训练语料集合 $T$，通过 Word2vec 训练语料库模型，输出语料模型 $M$，根据 $M$ 获得词向量 $vec_i$，对词向量取平均值后利用余弦相似度计算问题与答案的相似度，具体公式为：

$$S(A,B)=cos(\theta)=\frac{A \cdot B}{||A||\,||B||} \qquad \text{公式(9)}$$

在公式(9)中，$A$ 为问题文本词向量的均值，$B$ 为答案文本词向量的均值。每个答案文本的向量依次与问题文本的向量进行计算得到该问题与每个答案的语义相似度 $S(A,B)$。

Step4：计算加权融合的答案有用性排序数值 $P$。笔者将 $P$ 定义为语义相似度与加权灰色关联度的和。为了避免某一方数值过大或过小引起结果偏差，因此将 $P$ 改为语义相似度与加权灰色关联度的加权求和。为了获得较合适的权重值，笔者首先将权重值都设置为0.5并获取排序结果，发现由于 $\gamma_i$ 较大，$S$ 较小导致的排序结果不理想。笔者以0.5为基础进行权重值的调整和优化，经过多次试验和调参后，最终发现数值偏差对试验结果的影响最小的权重设置，计算公式如下所示：

$$P=0.25\gamma_i+0.75S(A,B) \qquad \text{公式(10)}$$

## 4 实证研究

### 4.1 数据采集与预处理

携程旅游网经过10年的发展已经成为功能健全的专业性旅游服务类网站。携程问答社区以用户生成答案为主，具有问题丰富、答案质量高、用户群体多元等特点。因此，选取携程旅游网问答社区作为研究对象，验证笔者提出的答案排序方法的可行性和有用性。携程问答作为旅游类的社会化问答社区的特殊性在于，用户提出的问题与生成的答案往往是与某个目的地有关。然而由于社会化问答社区存在一定的社交属性，每个目的地下的问答对都有可能存在无效数据。因此，为了能够最大限度地获取连贯有效的数据集，规避某些孤立数据造成答案有用性计算的偏移，也为了保证用户群体的多样性，笔者选择以多个目的地为检索词，随机采集每个目的地下一个问题及问题下的所有答案进行实证研究。首先利用八爪鱼数据采集软件以“杭州、上海、青岛、青海、武汉、长沙、三亚”为目的地检索词随机爬取页面下的一个问题的问答对文本内容、回答者粉丝数量、问题和答案的发布时间、用户的点赞数量、答案被回复的数量，答案的排列顺序。形成初始数据集总计924条，由于存在用户注销账户、用户重复回答和答案与问题无关等现象，导致出现无法获取相关数据、数据冗余和数据无效等问题，因此删除无效数据后剩余703条有效数据。问题和答案数量的相关实验数据如表2所示：



表2 问题和答案数量的相关实验数据

| 目的地 | 问题 | 答案数量 |
| --- | --- | --- |
| 长沙 | 作为小吃之都，长沙最地道的小吃在哪里？ | 263 |
| 杭州 | 最近计划去杭州踏青，初步定在4月初的一个周末，周六早上动车到杭州，然后一天游览西湖周边。我的规划是从断桥开始一路沿白堤步行，然后在苏堤北岸的曲院风荷游览后走到杨公堤，一路游览郭庄到茅家埠，再坐公交到龙井，午后再从龙井出发坐公交到苏堤南边的花港观鱼，看看雷峰塔和静寺，不知这样的规划是否合理呢？或者各位还有没有更好的建议？杭州西湖一天这样游合理吗？ | 113 |
| 青岛 | 我现居青岛 感觉青岛如果没有海也只是个一般的城市 但我又好齐三亚这个城市是怎么样的 有没有能给我比较比较这俩？三亚与青岛相比 哪个旅游价值更高？ | 50 |
| 青海 | 求西宁，青海湖，茶卡盐湖三日游路线 | 45 |
| 上海 | 下月准备带孩子去上海迪斯尼乐园，求大家推荐一下住宿的酒店和攻略 | 62 |
| 三亚 | 三亚都有那些值得去的景点？ | 113 |
| 武汉 | 我两天后要去武汉，武汉有哪些好玩的地方？ | 57 |

### 4.2 有用性排序方法应用

实验数据经过 python3.6 预处理后，分别按照前文的方法进行指标的量化处理，通过熵权法计算后获得问答社区的有用性指标权重分布结果见表3。

随机选取以“杭州”为目的地检索的随机问题“最近计划去杭州踏青，初步定在4月初的一个周末，周六早上动车到杭州，然后一天游览西湖周边。我的规划是从断桥开始一路沿白堤步行，然后在苏堤北岸的曲院风荷游览后走到杨公堤，一路游览郭庄到茅家埠，再坐公交到龙井，午后再从龙井出发坐公交到苏堤南边

表3 社会化问答社区答案有用性排序指标权重矩阵

| 问题主题 | 答案获赞 | 答案回复 | 图片数量 | 回答者权威 | 答案时效性 | 文本长度 | 回答者获赞 | 情感分析值 | 属性特征词 |
|---|---|---|---|---|---|---|---|---|---|
| 长沙 | 0.056 967 | 0.148 407 | 0.113 295 | 0.139 029 | 0.147 694 | 0.022 566 | 0.135 484 | 0.011 612 | 0.224 946 |
| 杭州 | 0.077 678 | 0.158 948 | 0.214 337 | 0.177 914 | 0.047 115 | 0.055 946 | 0.149 884 | 0.012 807 | 0.105 373 |
| 丽江 | 0.102 885 | 0.164 675 | 0.142 781 | 0.155 602 | 0.125 624 | 0.074 170 | 0.126 566 | 0.046 448 | 0.061 250 |
| 青岛 | 0.058 879 | 0.240 916 | 0.268 893 | 0.158 590 | 0.036 408 | 0.106 947 | 0.091 131 | 0.008 152 | 0.030 085 |
| 青海 | 0.091 122 | 0.073 482 | 0.176 759 | 0.244 718 | 0.110 197 | 0.062 965 | 0.139 360 | 0.023 406 | 0.077 992 |
| 上海 | 0.053 521 | 0.202 192 | 0.190 507 | 0.233 550 | 0.129 677 | 0.049 410 | 0.083 995 | 0.013 627 | 0.043 522 |
| 三亚 | 0.055 393 | 0.138 173 | 0.110 967 | 0.183 749 | 0.327 965 | 0.083 767 | 0.063 125 | 0.005 715 | 0.031 145 |
| 武汉 | 0.083 583 | 0.175 209 | 0.152 065 | 0.238 506 | 0.095 373 | 0.054 060 | 0.125 770 | 0.027 465 | 0.047 969 |

的花港观鱼,看看雷峰塔和静寺,不知这样的规划是否合理呢? 或者各位还有没有更好的建议? 杭州西湖一天这样游合理吗?”下的答案为例,运用加权灰色关联分析法融合 Word2vec 方法进行答案有用性排序。由于受到篇幅限制,仅选取第 4 个答案文本进行演示。

(1)各答案的灰色关联度值计算。首先,选取参考序列和比较序列。参考序列的各指标采用前述表 1 中指标的量化方法进行量化。例如:第 4 条答案的各项指标被量化后为{72,6,5,1084,2,464,119,1,13}。选择每个问题下答案的各个指标的最优指标数值作为参考序列,以杭州为主题的答案的参考序列为:

$Y$ = {0.679856,0.015108,0.010791,2.153957,0.021583,1.001439,5.082734,0.002158,0.032374}

其次,选择均值化为指标进行无量纲化处理后的结果为:

$X$ = {0.366931,0.030578,0.025481,5.524349,0.010193,2.364666,0.606455}

第三步,利用公式计算各个点之间的关联系数 Z:

$Z$ = {0.882035,0.999566,0.997083,0.626452,0.994172,0.961422,0.706199}

最后,根据公式(8)求出杭州为主题的问题下第 4 条答案与参考序列 Y 的加权灰色关联度 $\gamma_i$ = 0.097 451。

(2)答案与问题的语义相似度计算。由于携程问答社区中用户生成的答案是由大量非结构化的口语、网络语言构成,所以,笔者选取最全面的中文维基百科的语料库,利用 python3.6 进行分词等处理后,经过多次实验,最终选择词向量的训练维度为 256,窗口为 5,进行 word2vec 模型训练。笔者选择 skip-gram 模型训练语料并将所有数据转化成词向量,然后对词向量取平均值后,利用向量的余弦夹角求出问题与答案的相似度。同样以第 4 个答案作为演示对象。经过分词、去停用词处理后,问题转化为词列表[‘计划’,‘杭州’,‘西湖’,‘断桥’,‘白堤’,‘苏堤’,‘曲院荷风’,‘杨公堤’,‘郭庄’,‘茅家埠’,‘龙井’,‘花港观鱼’,‘雷峰塔’,‘静寺’,‘规划’],答案按照同样步骤转化为词列表[‘雷峰塔’,‘苏堤’,‘花港观鱼’,‘北山路’,‘孤山路’,‘白堤’,‘断桥 ’,‘毛家埠’,‘龙井’,‘灶丰年间’,‘弄堂里’,‘湖滨商业街’‘热门’,‘餐厅’];运用 Word2vec 模型将问题和答案转化成词向量,问题 $A = [v_1, v_2, \cdots, v_{15}]$,答案 $B = [v_1, v_2, \cdots, v_{14}]$,然后分别取问题与答案的词向量的平均值 $v_a$,$v_b$,利用向量的余弦夹角值计算出问题与答案的语义相似度。根据公式(9)求得问题 A 与答案 B 的语义相似度 $S(A,B)$ = 0.813 627。

最后,将答案的加权灰色关联度与问答对之间的语义相似度融合,根据公式(10)获得以杭州为主题的问题下第 4 条答案的最终答案有用性排序数值 P 为 0.270 854 225。

由于文章篇幅有限,选取以杭州为主题的问题及前 5 条答案进行对比分析。具体分析内容见表 4。

### 4.3 结果分析与对比

人工排序方法可以最直观地体现用户需求,因此被认为是最佳排序结果[50]。为了验证答案有用性排序方法的意义,本研究选择人工排序的方法辅助进行实验结果的对比分析。根据百度指数人群画像 2021 年 2 月 26 日到 2021 年 3 月 28 日的分析结果,携程的用户群年龄主要分布在 20 – 39 岁之间。因此,笔者随机选取 20 名此年龄段且拥有丰富的携程使用经验的用户,将答案顺序随机打乱后,要求实验参与者阅读每个问题及答案并从自我感知的角度出发,基于每条答案的内容丰富度、个人信息需求的满足程度、答案的有用程度等方面人工为答案排序。排序结果的答案重合率计算,不仅可以得出排序方法之间的优异,还可以得出答案有用性排序结果与最佳排序结果的接近程度。因此,笔者将 20 份人工排序结果进行对比整合后,提取本研究答案排序结果与人工排序结果的前 10 条答案,针对前 10 条答案计算出答案的重合率,具体结果见表 6。

表 4　答案有用性的排序结果

| 答案内容 | 有用性排序数值 | 有用性排序 | 携程原始排序 |
|---|---|---|---|
| 西湖最大的特点就是景多,所以玩法也特别多,不同的人可以玩出不同的感觉,从 1 天到 3 天,都能有不一样的游玩体验。具体的可以看看高德地图里的“杭州西湖一键智慧游”,里面有很多路线的推荐,从根据时间的 1 - 3 天路线,还有根据主题的茶文化路线、浪漫路线、骑行路线等等。像一天的话可以选择“精华一日游路线”,西湖著名的几个景点都包含在内了,飞来峰、灵隐寺、岳王庙、苏堤、三潭印月、白堤、音乐喷泉,有山有水,先去飞来峰看看石窟,出来去对面灵隐寺,然后到西湖周边参观下岳王庙、坐船到湖里看看三潭印月这些,等到了晚上去看看喷泉和夜景,一共 23.4 公里的行程。如果只想在西湖边上转转,可以选择“醉美十景”路线,看看断桥残雪、平湖秋月、曲院风荷、花港观鱼、雷峰夕照这些最经典的西湖十景,走累了还可以选择坐观光的电瓶车。如果喜欢骑行,还可以选择“沿湖骑行”路线,边骑车边看景,也是别有一番风味的。另外特别方便的一点是,每个景点都会有语音解说,都不用请导游了,而且还标注了景点之间直接可以乘坐的公交车或者步行导航的路线,真是有了这个就可以说走就走 | 0.295 959 | 1 | 4 |
| 以后去杭州可以这样走:西湖十景,是西湖上十处特色风景,有苏堤春晓、曲院风荷、平湖秋月、断桥残雪、柳浪闻莺、花港观鱼、雷峰夕照、双峰插云、南屏晚钟、三潭印月,一处比一处美,就这样,走着、看着,这样就挺好 | 0.292 931 | 2 | 24 |
| 动车杭州下来,如果是城站,那直接游 2 路到雷峰塔,浏览完后顺路走到苏堤,然后从苏堤南端往北走,苏堤比较长,全程 2.6km 还是 3.6km 忘记了,以前每次都是从头到尾暴走,现在基本上是走一段拍拍照,然后就往回走,你可以走一段拍拍看看,走到花港观鱼这里,然后做电瓶车继续往北,到了北山路这里右转,沿着孤山路一直走,经过白堤可以到达断桥,断桥附近可以坐公交去毛家埠和龙井,午饭可以在那里解决,有灶丰年间龙井店,和弄堂里茅家埠店哦,都是杭州特色性价比比较高的热门餐厅哦。吃完后,可以坐公交去湖滨商业街,这样就节省时间和体力了,忘亲采纳哦 | 0.291 922 | 3 | 35 |
| 西湖旅游应该是一种休闲旅游,如果只是走马观花、到此一游的话,建议选择一两个景点看看就可以了。我在西湖游了两天,但是也仅仅沿西湖转了两圈而已,连九溪十八涧和灵隐寺都没来得及去。如果想对西湖各个景点有个大致的了解,可以先坐电瓶车环游一圈,40 元一人,要游一个多小时;还可以租辆单车,骑行一圈。西湖沿线骑自行车还是很舒服的,特别是苏堤和白堤,特别适合跑步和骑行。晚餐可以在柳浪闻莺附近的莲遇餐厅用餐,是杭帮菜,如果美团的话要提前一天预约。用餐环境非常好,价格也不算太贵。那里还可以住宿,但是不便宜。晚上可以看看印象西湖。时间充裕的话可以用一天时间走走九溪十八涧和灵隐寺,还可以在灵隐寺附近的安曼喝喝茶,住不起安曼的酒店,也可以享受一下安曼的环境啊! 如果专门去游西湖的话,建议住西湖附近。因为杭州的公交车很挤,出租车基本打不到 | 0.288 416 | 4 | 6 |
| 8:30 - 10:30:游览灵隐寺飞来峰景区,寻双峰插云 10:30 - 12:30:游船环游西湖,在湖中观苏堤春晓、平湖秋月、断桥残雪、柳浪闻莺,登三潭印月岛 12:30 - 13:30:午餐 13:30 - 14:30 看曲院风荷 14:30 - 15:30 杨堤景行(杨公堤——西湖新十景之一,就在行程附近,同样值得一去)15:30: - 16:30 花港观鱼 16:30 - 18:00:听南屏晚钟,观雷峰夕照 tips:雷峰夕照步行 380 米于净寺站乘坐——315/344 路,30 分钟左右的车程到胡雪岩故居公交车站,步行 600 米,开始你的河坊街夜游,小吃应有尽有可在此解决晚餐(本是三天两夜杭州攻略的) | 0.285 104 | 5 | 17 |

表 5　携程问答的排序结果

| 答案内容 | 有用性排序数值 | 有用性排序 | 携程原始排序 |
|---|---|---|---|
| 这个设计感觉是在暴走啊,一天下来会特别累,如果是我的话,我会这样走:动车杭州下来,如果是城站,那直接游 2 路到雷峰塔,浏览完后顺路走到苏堤,然后从苏堤南端往北走,苏堤比较长,全程 2.6km 还是 3.6km 忘记了,以前每次都是从头到尾暴走,现在基本上是走一段拍拍照,然后就往回走,你可以走一段拍拍看看,走到花港观鱼这里,然后做电瓶车继续往北,到了北山路这里右转,沿着孤山路一直走,经过白堤可以到达断桥,断桥附近可以坐公交去毛家埠和龙井,午饭可以在那里解决,有灶丰年间龙井店,和弄堂里茅家埠店哦,都是杭州特色性价比比较高的热门餐厅哦。吃完后,可以坐公交去湖滨商业街,这样感觉不会有暴走的感觉哦,双脚第二天还可以继续用,哈哈。希望对你有帮助 | 0.281 339 | 6 | 1 |
| 我是携程当地向导,您好! 杭州两天,第一天西湖景区灵隐,第二天西溪湿地宋城,第三天乌镇(西塘三个小时就逛完了,白天好玩,夜景没乌镇好,建议去最有代表性的古镇乌镇,乌镇分东栅西栅两个景区,东栅主要看老房子,西栅主景区夜景非常好,建议买连票 150,分开买会贵 70 块)第四天苏州,苏州就是看园林,看苏州文化,主要拙政园,其他的几个园林去不去都无所谓了,大同小异,希望能帮到你,谢谢 | 0.278 449 | 7 | 2 |
| 西湖一天怎么游都合理 | 0.160 240 | 56 | 3 |
| 西湖最大的特点就是景多,所以玩法也特别多,不同的人可以玩出不同的感觉,从 1 天到 3 天,都能有不一样的游玩体验。具体的可以看看高德地图里的“杭州西湖一键智慧游”,里面有很多路线的推荐,从根据时间的 1 - 3 天路线,还有根据主题的茶文化路线、浪漫路线、骑行路线等等。像一天的话可以选择“精华一日游路线”,西湖著名的几个景点都包含在内了,飞来峰、灵隐寺、岳王庙、苏堤、三潭印月、白堤、音乐喷泉,有山有水,先去飞来峰看看石窟,出来去对面灵隐寺,然后到西湖周边参观下岳王庙、坐船到湖里看看三潭印月这些,等到了晚上去看看喷泉和夜景,一共 23.4 公里的行程。如果只想在西湖边上转转,可以选择“醉美十景”路线,看看断桥残雪、平湖秋月、曲院风荷、花港观鱼、雷峰夕照这些最经典的西湖十景,走累了还可以选择坐观光的电瓶车。如果喜欢骑行,还可以选择“沿湖骑行”路线,边骑车边看景,也是别有一番风味的。另外特别方便的一点是,每个景点都会有语音解说,都不用请导游了,而且还标注了景点之间直接可以乘坐的公交车或者步行导航的路线,真是有了这个就可以说走就走 | 0.295 798 | 1 | 4 |
| 第一站断桥步行约 20 分钟到楼外,然后曲院风荷,苏堤春晓,花港观鱼,午饭去龙井山吃农家菜,喝上一杯龙井茶,看看龙井茶园再拍拍相片,太阳快落山的时候雷锋塔,看雷峰夕照,晚饭杭州酒家尝尝正宗的杭州菜,晚饭后步行三分钟就能到西湖的湖滨路,看看音乐喷泉及西湖夜景,最后走走杭州唯一的仿古建筑河坊街、南宋御街 | 0.276 841 | 9 | 5 |

表 6 人工排序与原始排序的答案重合率对比分析

| 目的地 | 长沙 | 杭州 | 青岛 | 青海 | 上海 | 三亚 | 武汉 |
|---|---|---|---|---|---|---|---|
| 本文 | 57% | 74.5% | 73% | 63% | 54% | 47.5% | 58.5% |
| 携程 | 20% | 50% | 55% | 35% | 50% | 55% | 45% |

表 6 表明在以长沙、杭州、青岛、青海、上海、武汉为目的地的问答对中，答案有用性排序结果与人工排序结果的重复率高于原携程排序；在以三亚为目的地的问答对中答案有用性排序结果与人工排序结果的重复率低于原携程排序。据此可得出，整体上本研究提出答案有用性排序方法与人工排序更相似，更能满足用户个性化的信息需求。

## 5 结语

笔者以携程问答社区为例，从用户需求和答案的有用性的角度出发，在前人研究的基础上综合考虑答案特征、回答者特征和答案的时效性构建答案排序指标体系，然后对各指标进行量化，利用熵权法客观分析答案内部所含有的信息熵值，确定各指标的权重。然后结合灰色关联分析法计算答案的加权灰色关联度，并通过 Word2vec 计算出问答对之间的文本相似度，最后结合权重计算出每条答案的最终得分，获得答案的排序结果。实验结果表明，将本研究与携程问答社区现有的答案排序结果相比，本研究排序靠前的答案，一般多为图文结合、答案内容丰富、用户点赞数高、用户评论的热度高、情感分析值高、属性特征词多。原有排序只考虑用户点赞数或用户回答时间，本研究考虑到的用户需求维度更多，更能满足用户的个性化信息需求。然而，本研究也存在一定的不足。对于各种类型的社会化问答社区来说，本文实验选取的数据规模也较为有限，且仅仅限制在携程问答这单一的在线问答社区。下一步，笔者将扩大社会化问答社区的研究范围及实验数据规模。



**参考文献：**

[1] 曲明成. 问答社区中的问题与答案推荐机制研究与实现[D]. 杭州:浙江大学,2010.
[2] 贾佳,宋恩梅,苏环. 社会化问答平台的答案质量评估－以“知乎”、“百度知道”为例[J]. 信息资源管理学报,2013,3(2):19－28.
[3] 李波,高文君,邱锡鹏. 基于语法分析和统计方法的答案排序模型[J]. 中文信息学报,2009,23(2):23－27,47.
[4] 孙振鹏. 面向问答社区意见选择类问题的答案融合技术研究[D]. 哈尔滨:哈尔滨工业大学,2012.
[5] 冯文政,唐杰. 融合深度匹配特征的答案选择模型[J]. 中文信息学报,2019,33(1):118－124.
[6] 谢正文,柏钧献,熊熙,等. 基于增强问题重要性表示的答案选择算法研究[J]. 四川大学学报(自然科学版),2020,57(1):66－72.
[7] MA W, LOU J, JI C, et al. ACLSTM: A novel method for CQA answer quality prediction based on question-answer joint learning[J]. CMC-computers materials & continua, 2021, 66(1): 179－193.
[8] JIWOON J, CROFT W, SOYEON B. A framework to predict the quality of answers with non-textual features[C]//The 29th annual international ACM SIGIR conference. New York: ACM, 2006.
[9] AGICHTEIN E, CASTILLO C, DONATO D, et al. Finding high-quality content in social media[C]//Proceedings of the international conference on Web search and web data mining. New York: ACM, 2008:183－194.
[10] 尹隽,彭艳红,刘鹏,等. 基于信息接受模型的在线评论有用性影响因素研究[J]. 江苏科技大学学报(自然科学版),2020,34(03):69－78.
[11] MUDAMBI S M, SCHUFF D. What makes a helpful online review? A study of customer reviews on Amazon. com[J]. MIS quarterly, 2010,34(1):185－200
[12] 谢陈博. 社交问答平台答案有用性评价影响因素研究[J]. 现代商贸工业,2019,40(10):56－59.
[13] 曾珍妮. 社会化问答社区答案有用性影响因素研究[D]. 大连:大连理工大学,2019.
[14] 王晨. 知乎网络问答社区中用户回答有用性的影响因素研究[D]. 哈尔滨:哈尔滨工业大学,2020.
[15] SHAH C, POMERANTZ J. Evaluating and predicting answer quality in community QA[C]//International ACM Sigir conference on research & development in information retrieval. New York: ACM, 2010.
[16] 李晨,巢文涵,陈小明,等. 中文社区问答中问题答案质量评价和预测[J]. 计算机科学,2011,38(6):230－236.
[17] ZHOU Z M, LAN M M, NIU Z Y, et al. Exploiting user profile information for answer ranking in CQA[C]//21st World Wide Web conference 2012. Lyon: ACM Press, 2012:767－774.
[18] 来社安,蔡中民. 基于相似度的问答社区问答质量评价方法[J]. 计算机应用与软件,2013,30(2):266－269.
[19] 易明,张婷婷. 大众性问答社区答案质量排序方法研究[J]. 数据分析与知识发现,2019,3(6):12－20.
[20] YANG L, QIU M H, GOTTIPATI S, et al. CQARank: jointly model topics and expertise in community question answering[C]//Proceedings of the 22nd ACM international conference on information and knowledge management. New York: ACM, 2013.
[21] 郭顺利,张向先,陶兴,等. 社会化问答社区用户生成答案质量自动化评价研究——以“知乎”为例[J]. 图书情报工作,2019,63(11):118－130.
[22] 易明,张婷婷,李梓奇. 多维特征下社会化问答社区答案排序研究[J]. 图书情报工作,2020,64(17):103－113.
[23] 施国良,陈旭,杜璐锋. 社会化问答网站答案认可度的影响因素研究——以知乎为例[J]. 现代情报,2016,36(6):41－45.
[24] 翟倩. 在线评论有用性排序模型研究[D]. 长春: 吉林大学,2017.
[25] 彭岚,周启海,邱江涛. 消费者在线评论有用性影响因素模型研究[J]. 计算机科学,2011,38(8):205－207,244.
[26] 林先杰. 在线评论有用性影响因素研究[D]. 广州:华南理工大学,2014.

[27] 张鹏飞. 面向在线问答社区的问题检索与回答抽取技术研究与实现[D]. 长沙:国防科技大学,2015.
[28] 田作辉. 非事实类问题的回答选取[D]. 哈尔滨:哈尔滨工业大学,2013.
[29] TOBA H,ZHAO Y M,ADRIANI M,et al. Discovering high quality answers in community question answering archives using a hierarchy of classifiers[J]. Information sciences,2014,261(5):101 – 115.
[30] 郭国庆,陈凯,何飞. 消费者在线评论可信度的影响因素研究[J]. 当代经济管理,2010,32(10):17 – 23.
[31] 彭丽徽,李贺,张艳丰,等. 基于品牌声誉感知差异的在线评论有用性影响因素实证研究[J]. 情报科学,2017,35(9):159 – 164.
[32] 李蕾,何大庆,章成志. 社会化问答研究综述[J]. 数据分析与知识发现,2018,2(7):1 – 12.
[33] 马松岳,许鑫. 基于评论情感分析的用户在线评价研究——以豆瓣网电影为例[J]. 图书情报工作,2016,60(10):95 – 102.
[34] 董丽丽,赵繁荣,张翔. 基于领域本体、情感词典的商品评论倾向性分析[J]. 计算机应用与软件,2014,31(12):104 – 108,194.
[35] 李进华,张婷婷. 社会化问答知识分享用户感知有用性影响因素研究——以知乎为例[J]. 现代情报,2018,38(4):20 – 28.
[36] 龚思兰,丁晟春,周夏伟,等. 在线商品评论信息可信度影响因素实证研究[J]. 情报杂志,2013,32(11):202 – 207,180.
[37] 胡鹏辉. 基于多模型的问答社区答案质量评价研究[D]. 南京:南京师范大学,2019.
[38] 叶超. 问答社区中的问题推荐方法研究[D]. 广州:华南理工大学,2019.
[39] LIU Y,HUANG X,AN A, et al. Modeling and predicting the helpfulness of online reviews[C]//Proceedings of the 8thIEEE International Conference on Web Mining. Piscatway:IEEE,2008:443 – 452.
[40] 郝媛媛,叶强,李一军. 基于影评数据的在线评论有用性影响因素研究[J]. 管理科学学报,2010,13(8):78 – 88,96.
[41] 程启月. 评测指标权重确定的结构熵权法[J]. 系统工程理论与实践,2010,30(7):1225 – 1228.
[42] 李帅,魏虹,倪细炉,等. 基于层次分析法和熵权法的宁夏城市人居环境质量评价[J]. 应用生态学报,2014,25(9):2700 – 2708.
[43] 信桂新,杨朝现,杨庆媛,等. 用熵权法和改进 TOPSIS 模型评价高标准基本农田建设后效应[J]. 农业工程学报,2017,33(1):238 – 249.
[44] 邱菀华. 管理决策与应用熵学[M]. 北京:机械工业出版社,2002.
[45] 刘思峰,党耀国,方志耕. 灰色系统理论及其应用[M]. 北京:科学出版社,1999:55 – 60.
[46] 邓聚龙. 灰色理论基础[M]. 武汉:华中科技大学出版社,2002:35 – 76
[47]李枫林,吴敏. 用户知识内化过程中信息需求及获取行为研究[J]. 情报理论与实践,2015,38(1):35 – 38,52.
[48]GAVILAN D,AVELLO M,GEMA M N. The influence of online ratings and reviews on hotel booking consideration[J]. Tourism management,2018,66(6):53 – 61.
[49] 冯花朴. 潜在信息需求转化为信息行为的机理分析[J]. 现代情报,2009,29(10):11 – 13.
[50] 程亚男,王宇. 基于语义情感相似度的问答社区答案排序研究[J]. 情报科学,2018,36(8):72 – 76,83.

**作者贡献说明：**
郭顺利：提出研究思路，确定选题，修订论文；
步辉：提出研究框架，数据获取及分析，撰写论文。

## Research on the Sorting Method of Answer Usefulness in Social Q&A Community Integrating Word2vec and WGRA-Taking Ctrip Q&A as an Example

Guo Shunli　Bu Hui
School of Communication, Qufu Normal University, Rizhao 276800

Abstract: [**Purpose/significance**] In order to solve the diversified information needs of users and the problem of redundant and overloaded answers in the social Q&A community, this paper proposes an answer usefulness ranking method oriented to users' personalized needs ,assists users to efficiently filter and obtain useful answer knowledge. [**Method/process**] First, through literature research and expert consultation, an answer usefulness evaluation index system was constructed from the three dimensions of answer characteristics, answerer characteristics and answer timeliness; Then, it integrated the user's personalized needs from the semantic level, designed an answer usefulness ranking method that combined WGRA and Word2vec, and realized the answer ranking oriented to user needs. [**Result/conclusion**] Through comparative analysis of experimental results, it is found that compared with traditional ranking methods based on "likes" and "answer time", the answer usefulness ranking method designed in this paper has higher user satisfaction and is more able to satisfy users' personalized knowledge demands.

Keywords: user demand　answer usefulness　WGRA　Word2vec　social Q&A